As Harry mentioned before, we already know the log-likelihood functions for the Weibull distribution.

In order to find the maximum likelihood estimation of and when both of them are unknown, all we need to do now is maximizing these two log-likelihood functions. Since both parameters are unknown, we may need to use some numerical optimization methods.

In the R Stats package, we found a function called nlm, that is, non-linear minimization. We can use it to find the maximum.

Before showing the calculation results, we need to recap the Kaplan-Meier curve. Here’s the Kaplan-Meier curve for the survival time in the dataset “cancer”. The x-axis represents time, and y-axis represents the survival rate, that is, the proportion of the survivors. The red “+” symbols mark where the observations are censored, and the black curve indicates the change of the survival rate over time. Besides, Kaplan-Meier curve does reflect censoring, so when we consider censoring and calculate the MLE of and , the corresponding Weibull distribution curve is supposed to fit the Kaplan-Meier curve.

Using the function nlm, we can calculate that the MLE of and are 1.314 and 0.00036 respectively. Now we are going to see the corresponding Weibull distribution curve. Let’s add this curve on the same plot to see what happens. Yes, it does fit!!! And obviously, it lies exactly in the 95% confidence interval. This is the case when we consider censoring. How about ignoring censoring? Here’s the result.

When ignoring censoring, the MLE of and are 1.467 and 0.0002 respectively. And we can notice that the corresponding Weibull distribution curve is to the left of the censored one. Actually, whatever dataset with only right censored data you use, the uncensored curve is supposed to lie below the censored one. Why?

(On whiteboard) Let’s see an example. Suppose we have three patients, whose survival time are 100, 200 and 300 respectively, and they are all right censored (+), which means that their real survival time should be larger. For example, … When ignoring censoring, we just use the survival time recorded to plot the survival curve. Like this… However, when considering censoring, the curve is like this. Since the observations are right censored and the actual value is larger (-->), the survival rate will decrease more slowly when we take right censoring in consideration. That’s why the uncensored curve is to the left of the censored one.

Then we focus on the green line, the uncensored curve. Since it directly uses the survival time recorded, it is supposed to fit the empirical distribution function, simply called ecdf. Ecdf is almost the same as cdf. The only difference is that ecdf is for the specific data, while cdf is for a distribution. Now we are going to add the ecdf on this plot. As was expected, the uncensored curve fits the ecdf well. (Summary…)

So, we see that Weibull distribution works very well when we do survival analysis using parametric methods, then we have Kevin to introduce the case of other distributions.